DOCUMENT RESUME

ED 351 357 TM 019 167

AUTHOR Frary, Robert B.; And Others
TITLE Testing and Grading Practices and Opinions in the Nineties: 1890s or 1990s?
SPONS AGENCY Virginia Polytechnic Inst. and State Univ., Blacksburg.
PUB DATE Apr 92
NOTE 31p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, April 21-23, 1992).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Classroom Techniques; Cluster Analysis; Educational Assessment; *Educational Attitudes; Educational Practices; Evaluation Methods; *Grading; Mathematics Teachers; Public School Teachers; Questionnaires; Science Teachers; Secondary Education; *Secondary School Teachers; State Surveys; *Student Evaluation; *Teacher Attitudes; Teaching Experience; *Test Use
IDENTIFIERS Teacher Surveys; *Virginia

ABSTRACT

A statewide survey of 536 randomly selected Virginia secondary school teachers of academic subjects explored practices and opinions concerning various aspects of classroom testing and grading. The main focus was on the use of preestablished percentage scales as a basis for evaluating students. The responses suggested that large proportions of teachers hold opinions and pursue practices contrary to what many measurement specialists would recommend. Cluster analysis identified a small group of teachers whose opinions were largely consistent with what measurement specialists would recommend. This group differed from five other cluster groups in that it contained a disproportionate number of mathematics and science teachers and its mean years of out-of-state experience was substantially greater than that of any other group. Opinions and practices characterizing each of the other groups were extremely diverse. Some groups held views that might be considered inconsistent or self-contradictory. These findings led to the recommendation that the study questionnaire be adapted for administration to groups undergoing inservice training in measurement to facilitate focusing instruction according to the characteristics of each group. Seven tables and one figure present study findings. Four appendixes contain the survey instrument, a follow-up request, and proposed instruments for other research and inservice training. (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
ROBERT B. FRARY
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED351357

Testing and Grading Practices and Opinions in the Nineties: 1890s or 1990s?

Robert B. Frary[1], Lawrence H. Cross, and Larry J. Weber
Virginia Polytechnic Institute and State University

Presented at the Annual Meeting of the
National Council on Measurement in Education
San Francisco, April, 1992

Abstract

A statewide survey of randomly selected secondary teachers of academic subjects explored practices and opinions concerning various aspects of classroom testing and grading. The main focus was on the use of preestablished percentage scales as a basis for evaluating students. The responses suggested that large proportions of teachers hold opinions and pursue practices contrary to what many measurement specialists would recommend. However, cluster analysis identified a small group whose opinions were largely consistent with what we would recommend as measurement specialists. This group differed from five other cluster groups in various ways. For example, it contained a disproportionate number of mathematics and science teachers, and its mean years of out-of-state experience was substantially greater than that of any other group. The opinions and practices characterizing each of the other groups were extremely diverse. Some groups held views that might be considered inconsistent or self-contradictory. These findings led to the recommendation that the study questionnaire be adapted for administration to groups undergoing inservice training in measurement to facilitate focusing instruction according to the characteristics of each group.

[1]Office of Measurement and Research Services, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061-0438. Tel.: (703)231-7285 Fax: (703)231-9307 E-mail: FRARY@VTVM1.BITNET

TM019167

BEST COPY AVAILABLE

Testing and Grading Practices and Opinions in the Nineties: 1890s or 1990s?

Robert B. Frary, Lawrence H. Cross, and Larry J. Weber
Virginia Polytechnic Institute and State University

In an interesting historical account of grading practices, Cureton (1971) reported that by the turn of the century "marking systems based on 100 points or 100 percent were pretty well entrenched in many quarters, schools and colleges as well as civil service programs" (p. 4). However, she noted that, during the first two decades of the century, percentage marking procedures were the target of a great deal of criticism and that most measurement specialists agreed that "grades were really just ranks" (p. 6). Indeed, Monroe (1917) remarked that "to define a grade of 'excellent' or 'A,' as 95 to 100 per cent is merely to substitute one descriptive term for another" (p. 418).

Although measurement specialists have long recognized that most school marking procedures, especially in secondary school academic subjects, produce only rankings, it appears that many educators today perceive percentage grading scales as representing absolute measures. Indeed many school districts adopt official percentage grading scales whereby letter grades are defined by percentage ranges. Unfortunately, many teachers feel that they must apply the same percentage ranges to the scores from their classroom tests. Of course, striving to write tests that will yield scores largely between, say, 60% or 70% and 100% correct serves to undermine the potential of the tests to provide reliable evaluations of the examinees. Numerous other problems could be cited, not the least of which is the practice of using factors other than student achievement in the determination of grades.

While it is widely perceived in measurement circles that teacher practices and beliefs concerning testing and grading are more in keeping with the 1890s than the 1990s, recent documentation of this phenomenon has not been very comprehensive. Case studies involving only a few teachers have provided some valuable insights but little generalizable information (e.g., Stiggins, Frisbie, & Griswold, 1989; Barnes, 1985). Broader studies have confirmed the tendency of teachers to include components other than student achievement in their grading and the prevalence of various other undesirable practices (e.g., Agnew, 1985; Terwilliger, 1987; Jones, 1990; Manke & Loyd, 1990; Wood, Bennett, & Wood, 1990; Manke & Loyd, 1991). However, these and other recent studies have had limitations such as small sample size, nonrandom selection of subjects or selection from within a single school or system, and lack of focus on a specific instructional setting (e.g., combining elementary and secondary teachers or teachers of vocational and academic subjects). Some of the studies cited did not distinguish between opinion and practice (e.g., a teacher adhering to a rigid percentage grading scale due to district requirements but of the opinion that such a practice is not desirable). None of the studies attempted to measure individual teachers' beliefs in order to identify areas of deficiency for possible remediation, though an excellent study by Green and Stager (1986) did relate teachers' attitudes with respect to the general desirability of classroom and standardized testing to their personal characteristics.

Purpose of the Inquiry

Consideration of the inappropriate measurement practices revealed in the research just reviewed suggested that secondary teachers of academic subjects may be (collectively) the most susceptible to their commission. Only a very small proportion of these teachers have taken academic courses in educational measurement. Indeed, no state currently requires such a course for certification to teach a secondary subject area. Moreover, secondary teachers of academic subjects are highly likely to be producing tests the scores from which provide no more than ranking information in a milieu where percent-correct scores are traditionally interpreted as measures of the percentage of some body knowledge that the students have learned. In contrast, in elementary schools and in secondary courses with strong performance elements, such as art, physical education and vocational courses, tests yielding criterion- or domain-referenced scores are much more feasible to produce and are much more prevalent. Another consideration that focuses attention on academic courses is the fact that use of non-achievement factors may be appropriate in some nonacademic areas, for example, using evidence of good or bad sportsmanship to influence grades in physical education and evaluating students' attitudes toward work in some vocational courses.

In view of this analysis, this study was initiated to document the extent to which misconceptions regarding testing and grading were present in a large and representative sample of secondary teachers of academic subjects. The ultimate purpose was to determine and characterize the need for remediation or training in measurement. It was anticipated that individual teachers would have varied opinions and practices, in many cases both good and bad. Based on prior studies, it was decided to investigate teachers' beliefs and practices as these related to the following broad questions:

To what extent do teachers interpret test scores as representing the percentage of knowledge that a student has learned?

How pervasive is the practice of assigning letter grades directly on the basis of percent-correct scores?

To what extent do teachers appreciate the need for relatively difficult tests if the ranking function is to be served optimally?

To what extent do teachers believe differences in percentage grading scales across school districts constitute real differences in standards?

To what extent do teachers endorse the use of factors other than achievement in determining course grades?

How do teachers determine the minimum passing score for a test?

In addition, teacher beliefs about the efficacy of multiple-choice tests seemed relevant, given recent negative commentary concerning this mode in professional circles and the popular press.

Adopting a Measurement Perspective for Remediation

With respect to the questions just listed, some value judgements are necessary if one is to consider the need for remediation and how to accomplish it. In the case of using nonachievement factors in the determination of secondary academic course grades, we would expect few if any measurement specialists to endorse this practice. Remediation then would involve informing teachers of the practical, legal, and ethical concerns associated with reporting, as measures of achievement, grades that in fact reflect other factors. The evaluation of answers to all of the other questions listed above depends on one's tendency to prefer norm-referenced[2] as opposed to criterion- or domain-referenced measurement practices for secondary academic courses. For example, a promoter of a criterion-referenced approach would not look askance at teachers reporting that they interpret test scores as representing the percentage of knowledge learned. Such a person would recommend remediation to foster this concept among teachers not agreeing with it, probably in conjunction with training in the development of tests whose scores could bear a criterion-referenced interpretation. In contrast, a someone opting for a norm-referenced approach would probably view beliefs or practices consistent with criterion-referenced testing negatively and would design remediation to inform teachers of appropriate norm-referenced testing practices.

As might be surmised from the initial paragraphs of this paper, we have adopted a norm-referenced approach to interpreting the responses of secondary teachers of academic subjects with respect to the questions listed above. This position was taken in consideration of the general circumstances surrounding the teaching of secondary academic subjects and the nature of the subjects themselves. Secondary academic courses typically cover complex and varied subject matter that does not lend itself easily to criterion- or domain-referenced measurement. A more critical factor, however, is the time commitment required of a teacher for the production of good domain- or criterion-referenced tests. Most secondary teachers of academic subjects teach five or six classes daily involving multiple subjects and over 150 students. The time they have for test preparation is likely to allow for no more than writing the best questions they can think of that cover the material they have taught. Finally, we contend that the primary purpose of testing in a secondary academic course is and should be for grade determination and that a norm-referenced approach should yield scores and grades in a way

---

[2]We are using the term norm-referenced in this context to refer to tests whose scores provide only ranking information about the examinees. Given this characteristic of the scores, the teacher must apply external information, effectively norms, in order to assign value labels, usually letter grades, to score ranges. These norms might be based on the performance of reference groups on similar tests, perhaps prior groups of students in the same course. Alternatively, the norms might be more informal, for example, the teacher's evaluation of the usual quality of work done by specific students known to the teacher within various score ranges.

that maximizes reliability and hence fairness. We believe that, under such circumstances, the best remediation approach is to inform teachers of the measurement implications of their current testing practices and provide advice as to how they can use a norm-referenced approach to measurement to improve what they do without requiring a significant additional investment of time.

Methods and Results

The questions above motivated the development of a 44-item questionnaire. The first 17 items asked for factual information, mostly about testing and grading practices. The remaining items solicited opinions about "testing and grading *in an academic course*" using a four-point agree/disagree response scale. Further it was specified that responses should be based on the assumption that "the tests are designed to measure knowledge of subject matter taught rather than mastery of specific objectives." A copy of the questionnaire is provided in Appendix A. It should be noted that truly extraneous factors that might be used in determining course grades were covered as opinion items 36-39 rather than asking the responders to admit directly that they used them. In contrast, the use of legitimate grading factors was covered by items 14-17, which requested responses reflecting actual extent of use. Also, in item 6 teachers could report whether they used percentage grading scales (a common district requirement) but in various other items could give their opinions regarding this practice.

The population sampled for the survey was all secondary teachers of academic subjects employed in Virginia public schools in the spring of 1991. Permission was obtained to use the Master Personnel File maintained by the Virginia Department of Education to identify all secondary teachers having at least a 50% teaching assignment in any combination of English, mathematics, science, social studies, or foreign languages. From the 15,807 such teachers, a random sample of 800 was drawn. The questionnaires were mailed individually to the teachers at their schools with a postpaid return envelope and a cover letter promising confidentiality of individual responses. A clearly visible three-digit responder identification number was on each questionnaire, and two additional mailings were sent to nonresponders at three-week intervals. The final mailing also contained a postcard that could be returned in lieu of the questionnaire. Responders could check various possible reasons for being unwilling to respond (see Appendix B) and were asked to provide a telephone number if willing to be interviewed briefly. (It was hoped that such interviews might help to characterize possible nonresponse bias.)

Completed questionnaires were received from 539 teachers for a response rate of 67%. Three were discarded because the responders were no longer teaching secondary school academic subjects. Waves of responses associated with the initial and repeat mailings were processed separately. Chi-square tests across the responses to each item and the three response waves resulted in only one probability of less than .05 (for item 41). This outcome was attributed to chance. Only 21 postcards were returned. The postcard responses (some multiple) were as follows: 13 objected to spending the time required to complete the questionnaire, five objected to the lack of anonymity, two were skeptical about the value of

the research, one considered the survey an invasion of privacy, one was bothered by the machine-readable response format, and two gave no reason. Only three provided telephone numbers. They were not called.

The three-digit identifiers were used to merge responses with personal data from the Master Personnel File. This process added to each responder's record the following information: date of birth, sex, ethnicity, year(s) degree(s) awarded, location(s) of school(s) granting degree(s), total years of teaching experience, teaching experience in Virginia, teaching experience in the present school district, and certification status with respect to each teaching assignment.

Table 1 lists statistics for the sample and corresponding population parameters. The district types listed in Table 1 resulted from an ad hoc classification of the responders' school districts. Appropriate chi-square and *t*-tests comparing the sample statistics to the population parameters yielded no probabilities of less than .05. The certification levels for the sample and population were not compared; they exceeded 98% for all subject areas except foreign languages, for which the certification level was about 95% for the sample and the population.

Table 2 provides response statistics for the 44 questionnaire items (numbered 1-21 and 30-52 on the actual questionnaire). One might be tempted to note the majority positions on various items and characterize a "typical responder." However, for these data, simple crosstabulations revealed that such an individual could hardly be said to exist. Over the latter 27 opinion items, fewer than 15% of the responders agreed with the majority position on as many as 20 items. This finding led to a factor analysis of the opinion responses. A correlation matrix was constructed using pairwise deletion in the case of missing responses. A principal components extraction of roots yielded eight eigenvalues greater than unity. However, a varimax rotation of seven factors (accounting for 53% of total variance) yielded the most interpretable solution. Table 3 gives the loadings with absolute values greater than .50 on these factors. The factors were interpreted as representing the extent to which (contrary to what we would recommend) the responders agreed that:

1. Tests difficult enough to maximize ranking effectiveness are (nevertheless) undesirable.
2. Districtwide percentage grading scales are generally desirable and effective.
3. A percent-correct score reveals the absolute amount of a student's knowledge.
4. Minimum passing scores for tests should be set at a fixed percentage of correct answers.
5. Extraneous factors (e.g., effort, conduct) should influence course grades.
6. Multiple-choice tests are undesirable.
7. Difficult tests are pedagogically unsound.

Factor-related scores were determined by averaging the responses to the items defining each factor. Responses to items 44 and 45 were reversed within scale 2, as were all items of scales 1 and 5. This was done so that, uniformly across the scales, a low score (indicating agreement with the statements above) would represent an opinion *contrary* to our recommen-

dations. In the case of omissions within a scale, the average value over the responses present was used. Because of failure to answer the questions on the back of the questionnaire (the large "OVER PLEASE" notwithstanding), scores on scales 1 through 5 and 7 are missing for 17 responders. Table 4 contains statistics for the factor-related scale scores.

Cluster analysis was undertaken to identify groups of responders with relatively homogeneous scale scores based on the opinion items. Only the first five scale scores and the 519 responders with all five of these scores were used in this process. Scale 6 was judged to be only peripheral to the major concerns of the study, while scale 7 apparently represented some belief about test difficulty different from that of direct concern within the study. The goal of the cluster analysis was assignment of all responders to a small number of cluster groups of meaningful size. A nonhierarchical method was judged best for this purpose (Lorr, 1983, p. 20). Procedure FASTCLUS of the Statistical Analysis System (SAS Institute, Inc., 1989) was employed to produce four through eight cluster groups first using standardized and then nonstandardized scale scores and employing varying FASTCLUS specifications (e.g., full vs. partial seed replacement). Under each condition, the pseudo-*F* statistics continued to increase as the number of cluster groups increased. However, for many runs and in all runs producing more than six groups, the result was one or more groups with fewer than 30 members. Also, the pseudo-*F* statistics for analyses with more than six groups were only slightly larger than for those with six groups, and all pseudo-*F*s for six groups were of similar size. Accordingly, the analysis producing six groups with the largest minimum group size was adopted. It used nonstandardized scale scores and the program defaults to produce a minimum group size of 48. Table 5 gives the scale score means and standard deviations for the six groups so constituted. Figure 1 graphs the scale score means for the six groups to display each group's profile.

Responses to items 1 through 18, scale scores 6 and 7 (not used in clustering), and all personal variables of the study except certification status were analyzed with respect to cluster group membership. Two-way chi-square analyses compared group membership with all dichotomous and nominal variables, namely, items 1-3, item 6, item 7, sex, ethnicity, degree level (bachelor's vs. master's and higher), location of most recent degree (instate vs. out-of-state) and district type (as listed in Table 1). One-way analyses of variance were performed across cluster groups for all other variables (which were at least ordinal), namely, items 4-5, items 8-17, scale scores 6 and 7, age, years since obtaining bachelor's and master's degrees, and years of teaching experience outside of the present district, outside of Virginia, and in total. This process involved 30 tests of significance, of which 14 yielded probabilities of less than .05. Table 6 (group membership vs. age, and years of teaching experience) and Table 7 (group membership vs. major teaching assignment) present what are perhaps the most cogent of these outcomes.

Other significant findings were as follows:

> Thirty-six of the 44 Group 1 members (82%) answering item 3 considered their major teaching assignment "college preparatory" compared to only 55% of the 466 other

teachers answering this question. This disproportionality largely accounted for a chi-square with a probability of .015.

Members of Group 2 averaged the largest proportion of *A*s and Group 5 the smallest (item 4, $p=.032$).

Disproportionately large numbers of teachers in Groups 1, 2 and 4 chose option 3 of item 6, while the opposite was true for Groups 3 and 6 ($p<.001$). These choices were consistent with the mean scores for Groups 1 through 4 on scale 4 concerning flexibility in the determination of passing points (see Table 5 and Figure 1). On the same basis, one might have expected Group 6 to be somewhat inclined to choose option 3. Actually, only 10% of this group chose option 3 compared to 19% for the total sample, more in line with their tendency to believe that percent-correct scores reveal absolute levels of knowledge (scale 3). Group 4, in contrast, also tended to agree on scale 3 but answered item 6 more in line with their scale 4 scores.

The pattern of responses to item 7 was similar to that for item 6 with disproportionately large numbers from Groups 1 and 4 choosing option 2 and the opposite for Groups 3 and 6 ($p=.002$). About the same proportion of Group 2, however, chose option 2 as for the entire sample. With respect to the comment about Group 6 in the preceding paragraph, it may be noted that their interpretation of the meaning of test scores (item 7) is consistent with their reported practice (item 6), while their tendency to endorse flexibility in the determination of passing points (scale 4) is contradictory to these positions.

Members of Group 6 indicated the heaviest use of short answer questions and Group 1 the least (item 8, $p=.045$).

Members of Group 2 indicated the heaviest use of essay questions and Group 1 the least (item 9, $p=.004$). This outcome is probably largely a reflection of the fact that Group 2 had the greatest proportion of English teachers (see Table 7).

Members of Group 6 indicated the heaviest use of multiple-choice questions and Group 3 the least (item 10, $p=.025$).

Members of Group 6 indicated the heaviest use of true-false questions and Group 1 the least (item 11, $p<.001$).

Members of Group 1 indicated the heaviest use of problem solving questions and Group 2 the least (item 12, $p=.019$). This outcome is probably mainly due to the large percentages of mathematics and science teachers in Group 1 (see Table 7).

Members of Group 4 indicated the greatest influence of class participation in determining end-of-term grades and Group 3 the least (item 17, $p=.044$).

Members of Group 4 indicated the greatest agreement that difficult tests are pedagogically undesirable and Group 5 the least (scale 7, $p=.022$). In contrast, these two groups had almost equal mean scores on scale 1 (see Table 5), which also concerns test difficulty. However, this outcome is not inconsistent with the correlation of only .04 between scores on these two scales (see Table 4) and supports the earlier suggestion that scale 7 pertains to an aspect of test difficulty distinct from that of scale 1.

Discussion

It would appear that the responders are highly representative of the population insofar as it can be characterized by available information. Moreover, this information is quite comprehensive. Nevertheless, there is the possibility that nonresponse bias exists. Based on the limited postcard returns, it seems likely that real or perceived time pressure was largely responsible for nonreturns. It is difficult to make conjectures about how this phenomenon might have affected responses differentially. In any case, with a 67% return rate, a focused effect would have had to be very strong to cause a substantial change in the tentative conclusions which follow.

Given the results of the factor analysis and cluster analysis, Table 2 must be interpreted cautiously. Any generalization one might make based on this table is likely not to apply to one or more well-defined subgroups of teachers. Nevertheless, a few broad outcomes seem worth mentioning:

As expected, very few teachers of secondary school academic subjects reported using a criterion-referenced approach to testing (8% chose option 1 of item 6).

While a substantial minority of teachers reported the belief that their test scores provided only ranking information (41% chose option 2 of item 7), even more reported use of percent-correct scores in conjunction with an apparently domain-referenced interpretation (46% chose option 2 of item 6 *and* option 1 of item 7). This outcome highlights the need for training teachers to develop and use norm-referenced tests in a manner consistent with good measurement practice (unless one believes that most of the 46% were producing tests whose scores could bear a domain-referenced interpretation).

Teachers (at least collectively) did seem to be providing an adequate variety of testing modes for their students as evidenced by their responses to items 8 through 13 (see Table 2).

Test and quiz scores were reported (commendably) as the most dominant factors in grade determination (item 14). However, many measurement specialists might be uneasy with the large proportions of teachers who reported daily homework and class participation as "important" factors (items 16 and 17).

Large majorities agreed or tended to agree with the use of gain, ability, and effort in determining grades (items 36, 37, 38). Even behavior (item 39) received a 31% endorsement. Some informal feedback indicates that item 37, concerning the use of gain (endorsed by 85%), may have been misleading to some responders. Although it referred to "how much a student *gains*" (italics in original), some responders may not have interpreted this to mean gain as distinct from final achievement level. A rewording of this item seems advisable (see **Recommendations for Further Research**).

Several interesting outcomes are evident from Table 4. The extremely low intercorrelations reported there attest to the orthogonality of the underlying constructs. In turn, this orthogonality attests to the prevalence of contradictory beliefs. For example, the correlation of only .31 between scales 2 and 3 allows for many responders to believe on the one hand that district-wide percentage grading scales are desirable while, on the other hand rejecting the idea that percent-correct scores reflect the absolute level of a student's knowledge. All but one of the mean scale values are less than the scale midpoint of 2.5. Thus, however idiosyncratically opinions may be distributed, there is widespread disagreement with what we and many other measurement specialists would recommend. The mean of 2.05 on scale 6, indicating widespread agreement that multiple-choice tests are undesirable, is especially unsettling, given that these teachers reported fairly extensive use of this mode. One can only wonder about the effect of using multiple-choice tests "frequently" or "always" (52% of responses to item 10) while having reservations about their efficacy. (The correlation across all responders between responses to item 10 and scores on scale 6 is only .30.)

Opinion differences among the six cluster groups, as shown in Table 5 and Figure 1, permit them to be characterized as follows:

Group 1: The smallest of the six groups (N=48) is characterized by disagreement on all five scales, though only marginally so on scale 5 (use of extraneous factors). This group, therefore, unlike any other, appears to be largely in agreement with a norm-referenced approach to measurement.[3]

Group 2: This group (N=75) is most strongly characterized by being opposed to difficult tests (scale 1). At the same time they tend, at least minimally, to disagree with district-wide percent scales, doubt that percent-correct scores are absolute measures,

[3]The standard deviations for this group on scale 1 and scale 5 are as large as for the entire sample. This outcome attests to the fact that the pseudo-F statistic for the cluster analysis producing six groups was not at its maximum possible level. Production of additional groups would have split Group 1 into more homogeneous subgroups. This step was not taken, because such small groups would not have been useful for a major purpose of the study, namely, identification of groups (of practical size) with homogeneous opinions. While Group 1 opinions on scales 1 and 5 are quite varied, their opinions on the other scales are much less so. Further, the Group 1 mean on scale 1 is, by any reasonable standard, significantly above that of any other group (the standard error of this mean is .08).

and prefer flexible bases for determining passing points (scales 2, 3, and 4). Finally, they fairly strongly favor use of extraneous factors to determine course grades (scale 5). This group may well consist mostly of teachers who would be considered softhearted or who really do not like to use tests to determine the grades that they give.

Group 3: This group (N=83) is at the opposite end of scales 2-5 compared to Group 2. Thus, they could hardly be considered softhearted and, indeed, may well exemplify the stereotypical strict or hardnosed image of teachers. However, it is interesting that they reject the use of difficult tests almost as strongly as Group 2. Quite possibly, it is only by virtue of easy tests that reasonable proportions of their students can obtain acceptable grades, given their apparent strict adherence to fixed percentage scales and their strong rejection of the use of extraneous factors.

Group 4: Scales 4 and 5 strongly characterize this group (N=91). They are the strongest group in advocating a flexible approach to determining passing points and the strongest in advocating the use of extraneous factors to determine course grades. Simultaneously, they support districtwide percentage scales and the idea that percent-correct scores represent absolute measures, quite in contradiction to their flexible position on scale 4. Many of these teachers may be quite arbitrary or manipulative in their assignment of grades. Their relatively high acceptance level for difficult tests (scale 1) could certainly be accommodated if this conjecture is correct.

Group 5: Uncertainty appears to characterize this group (N=103). Most of their means are near the scale midpoints.

Group 6: Like Group 4, this largest group (N=119) is characterized by inconsistency. They tend to agree with district-wide percentage scales and view percent-correct scores as absolute measures while tending to reject fixed minimum passing points. However, they tend to favor easy tests and are, on the average, neutral about the use of extraneous factors to determine course grades. These latter two characteristics suggest an absence of the arbitrary or manipulative approach to grading that may characterize Group 4, whose profile is otherwise similar.

At this point, one might be tempted to wonder whether the groups just described exist in some real way or simply reflect statistical exploitation of the data. After all, while the scale scores are reasonably internally consistent (see Table 4), their validity could be questioned. In the same vein, any cluster analysis method yields clusters regardless of the validity or the reliability of its inputs. We submit that the 14 significant statistical tests (out of 30) reviewed in the preceding section challenge any serious doubts about the reality of the groups. Almost all of these outcomes are extremely consistent with the group characterizations just presented. For example, Group 2 (possibly softhearted or antitesting) reported the greatest proportion of *A*s (p=.032). Group 4 (possibly arbitrary or manipulative) reported the heaviest use of classroom participation in determination of course grades and Group 3 (strict or hardnosed) the

least ($p=.044$). Group 6, characterized as inconsistent, tended to support flexible determination of passing points (scale 4) while tending *not* to choose options 3 and 2 of items 6 and 7 respectively ($p<.001$ and $p=.002$). While such outcomes from within the questionnaire speak well for the authenticity of the clustering outcomes, it is information external to the questionnaire that more convincingly demonstrates the authenticity of the groups.

We as psychometricians would tend to favor Group 1 and have the greatest concern with respect to Group 4 (possibly arbitrary or manipulative). How compelling then that the mean years of out-of-district and out-of-state experience for these two groups are respectively the largest and smallest by large and highly significant margins (see Table 6). These differences are clearly not experience-based, as evidenced by the nonsignificant differences across groups for mean age and total years of teaching experience. This outcome could be attributed to simple provincialism, but we have no detailed conjectures to offer as to why teaching outside of Virginia or moving around within Virginia should engender what we perceive as a better approach to measurement. Certainly, there seems to be no obvious reason why the "worst" group should have the least external experience by such a large margin. Nevertheless, one would have to be extremely conservative to suggest that outcomes such as these could have arisen from chance factors affecting group formation.

Additional compelling evidence for the authenticity of the groups comes from their major teaching assignments (see Table 7). (This information was internal to the questionnaire but was mirrored almost perfectly by the assignment data in the Master Personnel File.) Mathematics and science teachers predominated in Group 1, while social studies and foreign language teachers were extremely underrepresented. Group 2 (possibly softhearted or antitesting) had a predominance of English teachers. Group 4 (possibly arbitrary or manipulative) had strong underrepresentation of English teachers. Group 6 (characterized by inconsistency) had a strong overrepresentation of social studies and foreign language teachers. As in the case of external teaching experience, it is difficult to conjecture mechanisms that may have yielded this breakdown, but the outcomes have the ring of plausibility to us, and we believe they could not reflect chance assignment to groups.

Perhaps as interesting as the variables yielding significant differences across the cluster groups are some variables that yielded no significant differences. Among these are sex, ethnicity, age, school level (middle/junior high vs. senior high), district type (as listed in Table 1), and location of the college granting the most recent degree (instate vs. out-of-state). This last finding may seem contrary to the findings concerning external experience but probably only reflects the fact that many teachers with out-of-state experience have obtained graduate degrees in Virginia.

It should be noted that other studies have found rather weak but statistically significant relationships between attitudes concerning testing and some of the dependent variables that did not differ significantly across the cluster groups of this study. However, these relationships were observed across all subjects in those studies, not across groups constituted on the basis of attitude profiles. This study did not consider overall relationships between the

dependent variables and attitudes because knowledge of these relationships would not have contributed to its purposes.

Recommendations for Further Research

Based on the apparently successful efforts of this study to define and constitute groups of teachers with distinct attitude profiles concerning testing and grading, investigation of the ramifications of this outcome is recommended. Appendix C contains a proposed questionnaire based on the items constituting the five scales. Some of the items have been revised slightly based on informal feedback. This version is designed for administration to varied large samples of teachers of secondary school academic subjects to validate its factor structure and test reproducibility of the cluster analysis results. In addition, the extent and character of regional variation in the scale score means could be established. For these purposes, the Office of Measurement and Research at Virginia Polytechnic Institute and State University (for address see cover sheet footnote) will make available copies of the machine-readable Appendix C questionnaire to qualified investigators at cost. Completed questionnaires may be returned for processing free-of-charge followed by return of data files to the participating investigators.

Appendix D contains an adapted version of the items constituting the opinion scales. It is designed for administration to groups of secondary school teachers of academic subjects about to undergo inservice training in classroom testing and grading. As may be noted in Appendix D, responses to the scale items is followed by production of individual scale scores and profiles based on the item groupings developed in this study. Groups of participants with similar profiles should then be formed, and instruction should be customized according to the score profiles of each group. For example, there need be little or no mention of the undesirability of using extraneous factors in determining course grades for teachers like those in Group 3 (rigid graders but favor easy tests). Discussion of the questionnaire items themselves and the profile characteristics of each group could be used in the course of instruction to focus attention on the issues of concern. (Note that the group numbers on this instrument differ from those of the study.) The effectiveness of the process just outlined should be evaluated in various ways, such as postinstruction questionnaires asking participants the extent to which specific prior beliefs were challenged and the extent to which they may have changed their opinions. The Appendix D instrument is copyrighted. Permission to reproduce it for the purposes just outlined will be granted to qualified investigators.

Acknowledgements

We wish to thank Dr. Gary T. Henry, Deputy Superintendent for Research, Policy Development and Information Systems, of the Virginia Department of Education for making available the information used in establishing the sample of teachers surveyed. The project was supported by a grant from the College of Education and by the Office of Measurement and Research Services at Virginia Polytechnic Institute and State University.

References

Agnew, E. (1985). *The grading policies and practices of high school teachers.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Barnes, S. (1985). A study of classroom pupil evaluation: The missing link in teacher education. *Journal of Teacher Education, 36*(4), 46-49.

Cureton, L.W. (1971). A history of grading practices. *Measurement in Education, 2*(4), 1-8.

Green, K. E., & Stager, S. F. (1986). Teacher attitudes towards testing. *Measurement and Evaluation in Counseling and Development, 19*, 141-150.

Jones, M. S. (1990). *Preservice teachers beliefs about effective grading practices.* Paper presented at the annual meeting of the National Council on Measurement in Education, Boston, MA.

Lorr, M. (1983). *Cluster analysis for social scientists.* San Francisco: Jossey-Bass, Inc.

Manke, M., & Loyd, B. (1991). *A study of teachers' understanding of their grading practices.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Manke, M., & Loyd, B. (1990). *An investigation of non-achievement-related variables in teachers' grading practices.* Paper presented at the annual meeting of the American Educational Research Association, Boston, MA.

Monroe, W. S. (1917). *Educational tests and measurements.* New York: Houghton Mifflin.

SAS Institute, Inc. (1989). *SAS/STAT user's guide.* Cary, NC: author.

Stiggins, R. J., Frisbie, D. A., & Griswold, P. A. (1989). Inside high school grading practices: Building a research agenda. *Educational Measurement: Issues and Practice, 8*(2), 5-14.

Terwilliger, J. (1987). *Classroom evaluation practices of secondary teachers in England and Minnesota.* Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, DC.

Wood, P., Bennett, T., & Wood, J. (1990). *Grading and evaluation practices and policies of school teachers.* Paper presented at the annual meeting of the National Council on Measurement in Education, Boston, MA.

Table 1

Comparison of Sample (N=536) and Population (N=15807)[a]

	Sample		Population	
Ethnicity:	N[b]	%	N[b]	%
Black	59	11.1	1949	12.3
White	462	86.8	13587	86.0
Other	11	2.1	266	1.7
Sex: Male	171	32.1	4774	30.2
Degree level: Master's or higher	206	39.1	5818	37.5
Source of most recent degree:				
From a Virginia school	310	58.8	9460	61.0
District type:				
Affluent urban/suburban	237	44.2	7569	47.9
Less affluent urban	109	20.3	2844	18.0
Appalachian	41	7.7	1499	9.2
Small city/town/rural	149	27.8	3895	24.6
Years of teaching experience:	Mean	S.D.	Mean	S.D.
Total	13.8	7.7	14.1	8.7
Outside of present district	3.2	4.9	3.1	4.7
Outside of Virginia	1.5	3.6	1.4	3.4
Age	42.7	10.4	42.9	11.3
Years since bachelor's	19.1	9.0	19.1	9.2
Years since master's	14.4	7.7	14.5	7.3
Percentage of teaching assignment devoted to:				
English	18.9	31.1	19.7	31.3
Mathematics	18.2	30.4	16.7	29.7
Science	16.4	29.4	15.0	28.6
Social studies	11.8	25.4	12.7	26.5
Foreign languages	5.7	19.7	7.0	21.2

[a]Appropriate chi-square and t-tests yielded no significant differences between the sample and the population ($p > .05$).

[b]Ns in table may not account for entire sample or population due to small numbers of missing data.

Table 2

Response Percentages for Questionnaire Items (N=536)[a]

Item					
1	School level:	Middle/junior high 35%	Senior high 65%		
2	Major teaching area:	English 28%	Mathematics 26%	Science 22.4%	
		Social Studies 16%	For. Language 8%		
3	Major area mainly college preparatory:		Yes 57%	No 43%	
4	Percentage of *A*s in major area:		Less than 5 21%	6 to 10 41%	
		11-20 21%	21-30 10%	More than 30 7%	
5	Percentage of *F*s in major area:		Less than 5 35%	6-10 30%	
		11-20 24%	21-30 8%	More than 30 4%	
6	Assignment of letter grades to test scores based on:				
	Number and level of specific learning objectives mastered				8%
	Fixed percentage scale				73%
	Various factors, e.g., item difficulty and knowledge of student effort/ability				19%
7	Interpretation of test scores:				
	Scores indicate percentage of subject matter learned				59%
	Scores provide a ranking of students on knowledge of material on test				41%

8-13	**Frequency of use of various kinds of test questions:**	**Never**	**Seldom**	**Occasionally**	**Frequently**	**Always**
8	Short answer	5%	12%	28%	45%	11%
9	Essay	22%	19%	21%	27%	11%
10	Multiple-choice	5%	16%	27%	38%	14%
11	True-false	19%	28%	34%	16%	3%
12	Math/science problems	40%	7%	12%	19%	22%
13	Performance	10%	20%	34%	31%	7%

14-17	**Extent to which various factors are used in major area grading:**	**Not used**	**Little influence**	**Important**	**Most influential**
14	Test/quiz scores	0%	1%	45%	54%
15	Projects, term papers	10%	9%	71%	11%
16	Daily homework	2%	22%	71%	5%
17	Class participation	10%	34%	51%	5%

18-21	**Opinions about testing and grading *in an academic course*:** **Item content**	**Agree**	**Tend to agree**	**Tend to disagree**	**Disagree**
18	Essay tests superior to mult.-choice	32%	44%	21%	4%
19	Multiple-choice tests superficial	13%	34%	46%	8%
20	Partial credit scoring desirable	44%	45%	8%	4%
21	Correction for guessing desirable	10%	25%	37%	29%

Table 2 contd.

30-45 Opinions about testing and grading *in an academic course*:

	Item content	Agree	Tend to agree	Tend to disagree	Disagree
30	%s translate directly to grades	39%	44%	12%	5%
31	A priori min. pass points desirable	32%	30%	22%	17%
32	Varying pass points undesirable	25%	26%	32%	18%
33	Pass point based on student perf.	17%	47%	27%	9%
34	Pass point based on item difficulty	16%	52%	23%	9%
35	Cut points based on score distrib.	9%	26%	43%	23%
36	Student ability should affect grade	22%	42%	23%	13%
37	Student's gain should affect grade	27%	58%	10%	4%
38	Student behav. should affect grade	10%	21%	29%	41%
39	Student effort should affect grade	19%	47%	23%	12%
40	District-wide % scale desirable	51%	28%	14%	7%
41	Higher standards where grading scales are more stringent	28%	36%	24%	12%
42	Lower GPAs in districts with more stringent grading scales	19%	30%	38%	13%
43	Stringent scales help in later work	24%	41%	26%	9%
44	Districtwide % scales ineffective since teachers adjust test difficulty	9%	31%	37%	24%
45	% scales meaningless due to lack of standardization	14%	49%	25%	12%

46-52 Opinions about testing and grading *in an academic course* based on assumption that *no school or district policy* dictates specific percentage ranges for grades:

	Item content	Agree	Tend to agree	Tend to disagree	Disagree
46	Harder tests permit more confident assignment of letter grades	10%	23%	46%	21%
47	Tests should contain items fewer than half can answer correctly	10%	24%	43%	23%
48	Passing with less than 50% hard to justify	38%	38%	19%	5%
49	Item with less than 30% right indicates something wrong in teaching/testing	26%	44%	20%	9%
50	Difficult tests pedagogically unsound	11%	41%	37%	11%
51	Hard questions desirable for better students even if some can't answer	5%	23%	48%	23%
52	Harder tests fairer to students	5%	19%	56%	21%

[a]Number of responders varies slightly across items due to sporadic omission of responses.

Table 3

Varimax Rotated Factor Matrix[a]

		Factor						
	Item Content	**1**	**2**	**3**	**4**	**5**	**6**	**7**
18	Essay tests superior to mult.-choice						74	
19	Multiple-choice tests superficial						74	
20	Partial credit scoring desirable						58	
21	Correction for guessing desirable[b]							
30	%s translate directly to grades			64				
31	A priori minimum pass points des.			68				
32	Varying pass points undesirable			58				
33	Pass point based on student perf.				78			
34	Pass point based on item difficulty				80			
35	Cut points based on score distrib.				57			
36	Student ability should affect grade					62		
37	Student's gain should affect grade					62		
38	Student behavior should affect grade					56		
39	Student effort should affect grade					78		
40	District-wide % scale desirable		59					
41	Higher standards where grading scales are more stringent		61					
42	Lower GPAs in districts with more stringent grading scales[b]							
43	Stringent scales help in later work		60					
44	Districtwide % scales ineffective since teachers adjust test difficulty		-67					
45	% scales meaningless due to lack of standardization		-57					
46	Harder tests permit more confident assignment of letter grades	64						
47	Tests should contain items fewer than half can answer correctly	74						
48	Passing with less than 50% hard to justify			58				
49	Item with less than 30% right shows something wrong in teaching/testing							78
50	Difficult tests pedagogically unsound							66
51	Hard questions desirable for better students even if some can't answer	68						
52	Harder tests fairer to students	69						

[a]Loadings greater than .50 shown, decimal points omitted.
[b]No loading greater than .50.

Table 4

Scale Score Statistics (*N*=519)

Scale	Extent to which responders agree	Items in Scale	Mean[a]	S. D.	Coef. α
1	Tests difficult enough to maximize ranking effectiveness are (nevertheless) undesirable	46[b],47[b],51[b],52[b]	2.15	.49	.70
2	Districtwide percentage grading scales are generally desirable and effective	40,41,43,44[b],45[b]	2.26	.59	.65
3	Percent-correct score reveals the absolute amount of a student's knowledge	30,31,32,48	2.10	.65	.60
4	Minimum passing scores for tests should be set at a fixed percentage of correct answers	33[b],34[b],35[b]	2.55	.68	.70
5	Extraneous factors (e.g., effort, ability) should influence course grades	36,37,38,39	2.37	.62	.62
6	Multiple choice tests are undesirable	18,19,20	2.05	.58	.55
7	Difficult tests are pedagogically unsound	49,50	2.30	.69	.39

Intercorrelations Among Scale Scores

Scale	1	2	3	4	5	6	7
1	-	.11	.08	.19	.13	.01	.04
2	.11	-	.31	.16	.08	.02	.01
3	.08	.31	-	.26	.13	.09	.08
4	.19	.16	.26	-	.39	.00	.15
5	.13	.08	.13	.39	-	.11	.14
6	.01	.02	.09	.00	.11	-	.08
7	.04	.01	.08	.15	.14	.08	-

[a]Based on 1=Agree, 2=Tend to agree, 3=Tend to disagree, 4=disagree.
[b]Scoring reversed prior to averaging across scale items.

Table 5

Scale Score Means[a] (and Standard Deviations) by Cluster Group

Scale	Extent to which responders agree	Grp 1 *N*=48	Grp 2 *N*=75	Grp 3 *N*=83	Grp 4 *N*=91	Grp 5 *N*=103	Grp 6 *N*=119
1	Difficult tests undesirable	2.86	1.63	1.76	2.61	2.50	1.83
		(.53)	(.36)	(.45)	(.43)	(.44)	(.39)
2	District %age scale desirable	2.95	2.67	1.80	2.05	2.45	2.01
		(.50)	(.60)	(.47)	(.47)	(.46)	(.37)
3	%s reveal absolute knowledge	3.04	2.68	1.70	1.99	2.05	1.75
		(.53)	(.48)	(.57)	(.47)	(.46)	(.43)
4	Minimum passing % fixed	3.03	2.82	1.65	3.13	2.17	2.73
		(.54)	(.50)	(.47)	(.46)	(.46)	(.37)
5	Include extraneous factors	2.54	2.08	3.04	1.86	2.39	2.44
		(.69)	(.52)	(.50)	(.45)	(.41)	(.45)

[a]Based on 1=Agree, 2=Tend to Agree, 3=Tend to disagree, 4=Disagree

Table 6

Age and Years of Teaching Experience Means by Cluster Group

Variable	Grp 1 *N*=48	Grp 2 *N*=75	Grp 3 *N*=83	Grp 4 *N*=91	Grp 5 *N*=103	Grp 6 *N*=119	Prob.
Age	44.8	44.5	42.9	42.4	41.1	42.8	.423
Total teaching experience	15.5	14.1	14.4	13.6	12.8	13.3	.403
Out-of-district experience	7.9	3.1	2.8	1.6	2.7	3.3	<.001
Out-of-state experience	4.0	1.3	1.3	.6[a]	1.5	1.4	<.001

[a]Not a typo

Table 7

Group Membership According to Major Teaching Assignment[a]

		Percentage of Teachers in Each Group					
Major Assignment	***N***	**Grp 1**	**Grp 2**	**Grp 3**	**Grp 4**	**Grp 5**	**Grp 6**
English	144	7	23	18	12	19	21
Mathematics	135	13	9	17	20	18	24
Science	117	15	14	15	20	24	14
Social Studies	82	2	12	11	20	22	33
Foreign Language	41	2	10	20	20	15	34
Total	519	9	14	16	18	20	23

		Percentage of Teachers with Each Assignment				
Group Number	***N***	**English**	**Mathe-matics**	**Science**	**Social Studies**	**Foreign Language**
1	48	23	35	35	4	2
2	75	44	16	21	13	5
3	83	31	28	20	11	10
4	91	19	30	25	18	9
5	103	26	23	27	17	6
6	119	25	27	13	23	12
Total	519	28	26	23	16	8

[a] $\chi^2=41.3$ for group by major assignment; with 20 d.f. $p=.003$.

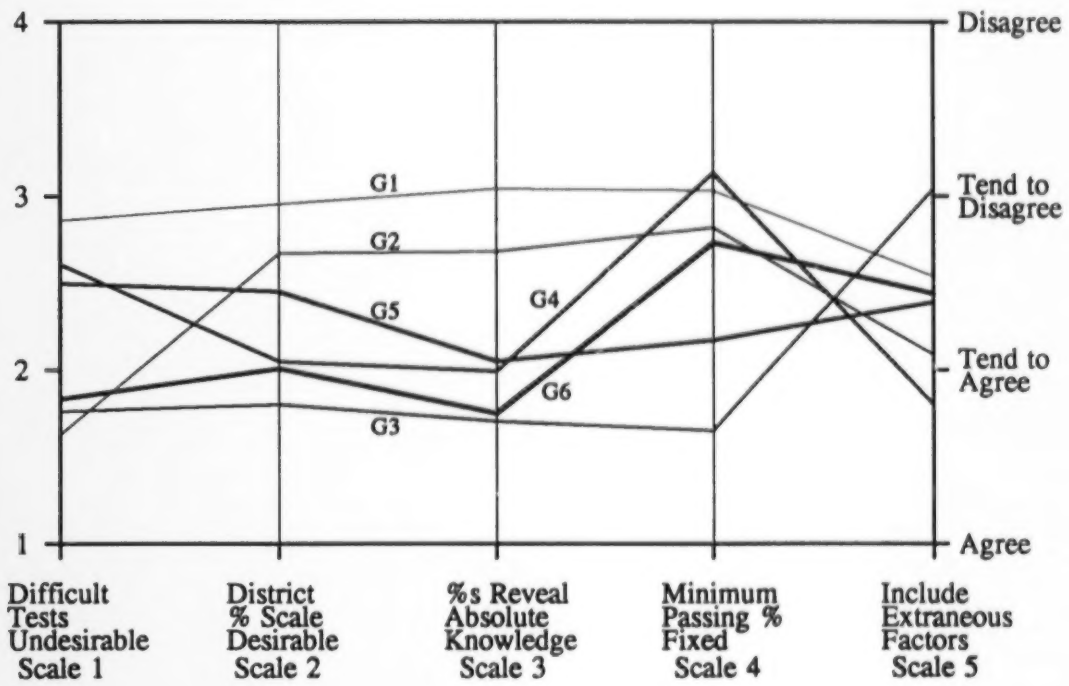


Figure 1. Group Proliles on Means of Clustering Variables

APPENDICES

A. Original Survey Instrument

B. Survey Postcard

C. Proposed Instrument for Further Research

D. Proposed Instrument for Inservice Training

VIRGINIA TECH

SURVEY OF TESTING AND GRADING PRACTICES

Your responses to this survey are requested to help in a study of testing and grading practices across different school systems. Please use a ***No. 2 Pencil*** to mark your responses in the answer column.

1. Your school level: 1) Middle or junior high school 2) High school
2. Your major teaching area (what you spend the most time teaching):
   1) English, including language arts, drama, journalism, etc. 2) Mathematics 3) Science 4) Social studies 5) Foreign language 6) Other (specify) _____
3. Are most of the classes you teach in your major teaching area considered "college preparatory"?
   1) Yes 2) No
4. In your major teaching area, about what percentage of your students earn *A*s?
   1) Less than 5% 2) 6% - 10% 3) 11% - 20% 4) 21% - 30% 5) More than 30%
5. In your major teaching area, about what percentage of your students earn *F*s?
   1) Less than 5% 2) 6% - 10% 3) 11% - 20% 4) 21% - 30% 5) More than 30%
6. Which of the following comes ***closest*** to describing how you generally assign letter grades to test scores?
   1) I use criterion-referenced tests targeted to specific learning objectives. Grades are based on the number or level of objectives mastered.
   2) I compute the percentage of correct answers and assign grades based on more-or-less fixed percentage ranges (e.g., 92% - 100% = *A*, 85% - 91% = *B*, etc.).
   3) I use various factors, such as how difficult the questions were, the scores earned by students whose work I know especially well, natural breaks in the score distribution, etc.
7. Which of the following comes ***closer*** to describing how you interpret the scores that students make on your tests?
   1) The scores indicate the percentage of the topic measured by the test that has been learned by each student.
   2) The scores provide a ranking of the students according to how much they know about the material covered in the test.

Questions 8 through 13 list various kinds of tests or test questions. Please indicate how often you use each type in your major teaching area according to the following scale:

1) Never 2) Seldom 3) Occasionally 4) Frequently 5) Always

8. Short answer
9. Essay
10. Multiple-choice/matching
11. True-false
12. Mathematics/science problems (showing solutions)
13. Performance (e.g., oral presentations, laboratory work)

Questions 14 through 17 list factors that are commonly used in determining report card grades. Please indicate the extent to which you use these factors in your major teaching area according to the following scale:

1) Not used at all in my grading
2) Used but of little influence
3) Important but not predominant
4) More influential than any other factor

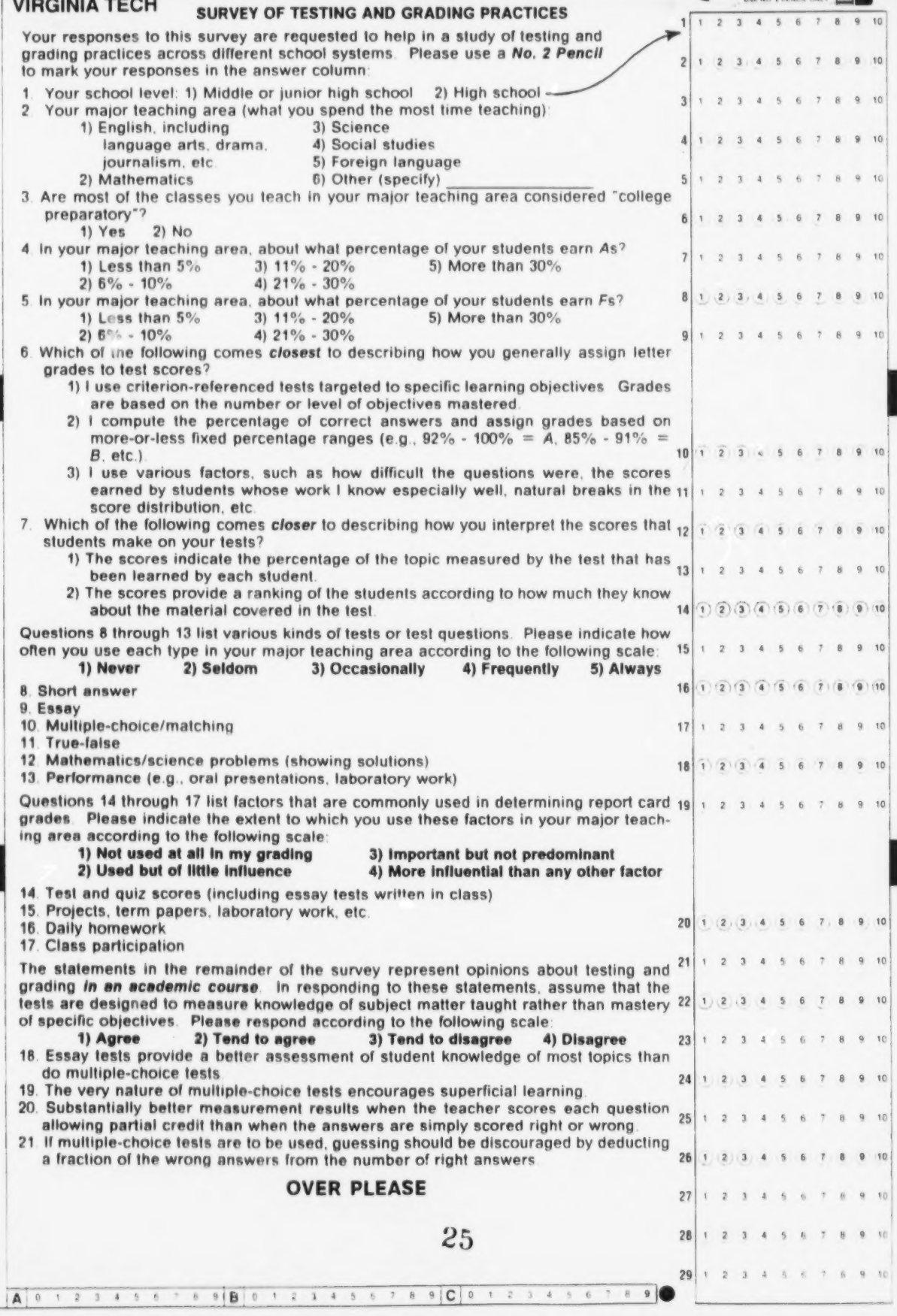
14. Test and quiz scores (including essay tests written in class)
15. Projects, term papers, laboratory work, etc.
16. Daily homework
17. Class participation

The statements in the remainder of the survey represent opinions about testing and grading ***in an academic course***. In responding to these statements, assume that the tests are designed to measure knowledge of subject matter taught rather than mastery of specific objectives. Please respond according to the following scale:

1) Agree 2) Tend to agree 3) Tend to disagree 4) Disagree

18. Essay tests provide a better assessment of student knowledge of most topics than do multiple-choice tests.
19. The very nature of multiple-choice tests encourages superficial learning.
20. Substantially better measurement results when the teacher scores each question allowing partial credit than when the answers are simply scored right or wrong.
21. If multiple-choice tests are to be used, guessing should be discouraged by deducting a fraction of the wrong answers from the number of right answers.

OVER PLEASE



Please continue to respond according to the same scale as before:

1) Agree **2) Tend to agree** **3) Tend to disagree** **4) Disagree**

30. Tests used to award grades should yield percent correct scores that translate directly into grades according to preestablished ranges.
31. A teacher should decide on the minimum passing score on a test before it is administered.
32. To decide that 60% correct is passing for one test and 70% for another test in the same class is arbitrary and capricious and should be avoided.
33. The minimum passing score on a test should be based at least in part on the scores earned by students of marginal ability who have been putting forth satisfactory effort.
34. The minimum passing score on a test should be based at least in part on how many of the test questions are easy enough for students of marginal ability to answer if they have been putting forth satisfactory effort.
35. Where to draw the line between grades of *A* and *B*, *B* and *C*, etc., should be decided only after looking at the distribution of scores for the class.
36. A student's ability should be taken into consideration in awarding the final grade.
37. The amount of knowledge a student ***gains*** over the instructional period should be taken into consideration in awarding the final grade.
38. Laudatory or disruptive classroom behavior should be considered in determining final grades.
39. An exceptionally low or high degree of effort should be recognized by adjustment of the final grade.
40. A district-wide percentage grading scale (e.g., 92%-100% = *A*, 85%-91% = *B*, etc.) is to be preferred over a qualitative one (e.g., *A* = Excellent, *B* = Good, etc.).
41. Students are likely to be held to higher standards in school districts that have more stringent grading scales (e.g., 95%-100% = *A* instead of 90%-100% = A).
42. Students are likely to have lower grade-point averages in school districts that have more stringent grading scales than students of equal ability in districts with less stringent scales.
43. Stringent grading scales better prepare students for the world of work, since a 60% correct standard is unacceptable in most work settings.
44. District-wide percentage grading scales are generally ineffective, since teachers can make their tests as easy or as hard as necessary to obtain the range of grades they desire to give.
45. Percentage grading scales are generally meaningless, since there is no standardization of what, say, 90% is a percentage of.

For the remaining statements, assume that there is ***no school or district policy*** that dictates specific percentage ranges for grades.

46. Letter grades can be assigned to test scores more confidently if the test is hard enough to make the scores range from, say, 30% to 90%, rather than a range of 60% to 100%, as would result from an easier test.
47. Tests used to award letter grades should contain at least a moderate proportion of questions that fewer than half of the students can answer correctly.
48. If a student answered fewer than 50% of the questions on a classroom test correctly, it would be hard to justify anything other than a failing grade.
49. If fewer than 30% of the students answer a test question correctly, there is something *wrong* in the teaching/learning/testing process.
50. Difficult classroom tests are pedagogically unsound, because low scores discourage students even if satisfactory grades result.
51. In order to challenge the better students, it is desirable to use test questions that many of the less able students may be unable to answer correctly no matter how hard they have studied.
52. A harder test with an average score of 60% is fairer to students than an easier test with an average score of 80% (for testing the same material).

Please return in the envelope provided to:

Office of Measurement and Research Services
Virginia Tech
2096 Derring Hall
Blacksburg, VA 24061-0438

Thanks for your help!

Please check below to indicate the reason(s) that you are not returning the questionnaire:

____ Don't wish to spend the time required.

____ Skeptical about the value of the research.

____ Bothered by machine-readable format (having to mark responses in answer column).

____ Consider the request to participate an invasion of privacy.

____ Object to the lack of complete anonymity (ID numbers on response sheets).

____ Other reason (please explain) _____

_____

_____

_____

May we phone you at home in the evening (7-9 pm) concerning your opinions about a small number of key issues on the questionnaire?

____ No

____ Yes, my home phone number is ( ) _____ - _____

Best days of week to call are _____

VIRGINIA TECH **Survey of Opinions About Testing and Grading**

The statements which follow represent opinions about testing and grading in an academic course. In responding to these statements, assume that the tests are designed to measure knowledge of subject matter taught rather than mastery of specific objectives. Also assume that a range of grades from A to F is expected.

Please use a **NO. 2 PENCIL** to mark your responses in the answer column, and be sure to match the question numbers with the answer column numbers. Respond according to the following scale:

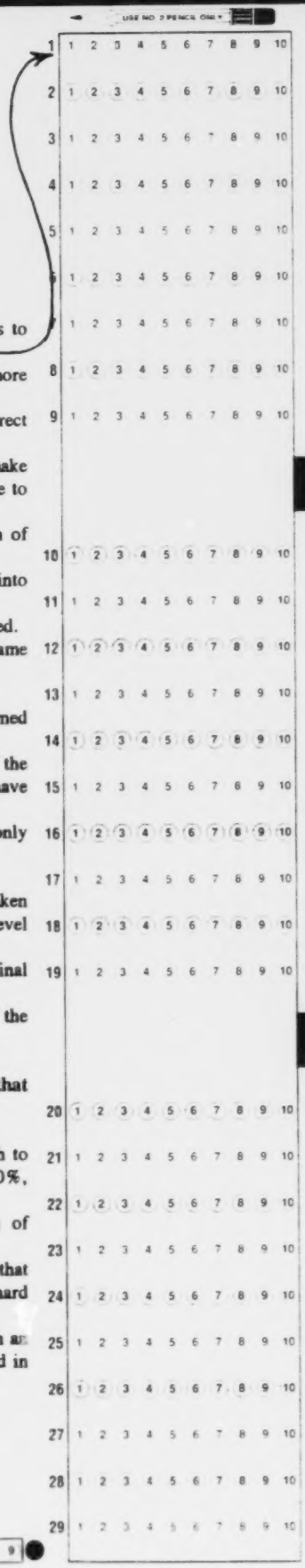
1) Agree 2) Tend to agree 3) Tend to disagree 4) Disagree

1. A district-wide percentage grading scale (e.g., 92%-100%=A, 85%-91%=B, etc.) is to be preferred over a qualitative one (e.g., A=Excellent, B=Good, etc.).
2. Students are likely to be held to higher standards in school districts that have more stringent grading scales (e.g., 95%-100% = A instead of 90%-100% = A).
3. Stringent grading scales better prepare students for the world of work, since a 60% correct standard is unacceptable in most work settings.
4. District-wide percentage grading scales are generally ineffective, since teachers can make their tests as easy or as hard as necessary to obtain the range of grades they desire to give.
5. Percentage grading scales are generally meaningless, since there is no standardization of what, say, 90% is a percentage of.
6. Tests used to award grades should yield percent correct scores that translate directly into grades according to preestablished ranges.
7. A teacher should decide on the minimum passing score on a test before it is administered.
8. To decide that 60% correct is passing for one test and 70% for another test in the same class is arbitrary and capricious and should be avoided.
9. A test score of less than 50% clearly represents a very low level of achievement.
10. The minimum passing score on a test should be based at least in part on the scores earned by students of marginal ability who have been putting forth satisfactory effort.
11. The minimum passing score on a test should be based at least in part on how many of the test questions are easy enough for students of marginal ability to answer if they have been putting forth satisfactory effort.
12. Where to draw the line between grades of A and B, B and C, etc., should be decided only after looking at the distribution of scores for the class.
13. A student's ability should be taken into consideration in awarding the final grade.
14. The amount of knowledge a student gains over the instructional period should be taken into consideration in awarding the final grade (in addition to the student's final level of achievement).
15. Laudatory or disruptive classroom behavior should be considered in determining final grades.
16. An exceptionally low or high degree of effort should be recognized by adjustment of the final grade.

For the remaining statements, assume that there is no school or district policy that dictates specific percentage ranges for grades.

17. Letter grades can be assigned to test scores more confidently if the test is hard enough to make the scores range from, say, 30% to 90%, rather than a range of 60% to 100%, as would result from an easier test.
18. Tests used to award letter grades should contain at least a moderate proportion of questions that fewer than half of the students can answer correctly.
19. In order to discriminate among the better students, it is desirable to use test questions that many of the less able students may be unable to answer correctly no matter how hard they have studied.
20. A harder test with an average score of 60% is fairer to students than an easier test with an average score of 80%, if the same numbers of A's, B's, C's, etc., will be assigned in either case.

Thanks for your responses!



30	1	2	3	4	5	6	7	8	9	10
31	1	2	3	4	5	6	7	8	9	10
32	1	2	3	4	5	6	7	8	9	10
33	1	2	3	4	5	6	7	8	9	10
34	1	2	3	4	5	6	7	8	9	10
35	1	2	3	4	5	6	7	8	9	10
36	1	2	3	4	5	6	7	8	9	10
37	1	2	3	4	5	6	7	8	9	10
38	1	2	3	4	5	6	7	8	9	10
39	1	2	3	4	5	6	7	8	9	10
40	1	2	3	4	5	6	7	8	9	10
41	1	2	3	4	5	6	7	8	9	10
42	1	2	3	4	5	6	7	8	9	10
43	1	2	3	4	5	6	7	8	9	10
44	1	2	3	4	5	6	7	8	9	10
45	1	2	3	4	5	6	7	8	9	10
46	1	2	3	4	5	6	7	8	9	10
47	1	2	3	4	5	6	7	8	9	10
48	1	2	3	4	5	6	7	8	9	10
49	1	2	3	4	5	6	7	8	9	10
50	1	2	3	4	5	6	7	8	9	10
51	1	2	3	4	5	6	7	8	9	10
52	1	2	3	4	5	6	7	8	9	10
53	1	2	3	4	5	6	7	8	9	10
54	1	2	3	4	5	6	7	8	9	10
55	1	2	3	4	5	6	7	8	9	10
56	1	2	3	4	5	6	7	8	9	10
57	1	2	3	4	5	6	7	8	9	10
58	1	2	3	4	5	6	7	8	9	10
59	1	2	3	4	5	6	7	8	9	10
60	1	2	3	4	5	6	7	8	9	10

Teacher Opinions on Testing and Grading

The statements which follow represent opinions about testing and grading in an academic course. In responding to these statements, assume that the tests are designed to measure **knowledge of subject matter taught rather than mastery of specific objectives**. Also, assume that the statements apply to classes where grades typically range from F to A. Decide how much you agree or disagree with each statement. Then write the number next to your choice in the box at the right of the question.

Note that the order of the numbers is reversed for some statements.

1. A district-wide percentage grading scale (e.g., 92%-100%=A, 85%-91%=B, etc.) is to be preferred over a qualitative one (e.g., A=Excellent, B=Good, etc.).
   Agree (1) Tend to agree (2) Tend to disagree (3) Disagree (4)
2. Students are likely to be held to higher standards in school districts that have more stringent grading scales (e.g., 95%-100% = A instead of 90%-100% = A).
   Agree (1) Tend to agree (2) Tend to disagree (3) Disagree (4)
3. District-wide percentage grading scales are generally ineffective, since teachers can make their tests as easy or as hard as necessary to obtain the range of grades they desire to give.
   Agree (4) Tend to agree (3) Tend to disagree (2) Disagree (1)
4. Percentage grading scales are generally meaningless, since there is no standardization of what, say, 90% is a percentage of.
   Agree (4) Tend to agree (3) Tend to disagree (2) Disagree (1)
5. Tests used to award grades should yield percent correct scores that translate directly into grades according to preestablished ranges.
   Agree (1) Tend to agree (2) Tend to disagree (3) Disagree (4)
6. A teacher should decide on the minimum passing score on a test before it is administered.
   Agree (1) Tend to agree (2) Tend to disagree (3) Disagree (4)
7. To decide that 60% correct is passing for one test and 70% for another test in the same class is arbitrary and capricious and should be avoided.
   Agree (1) Tend to agree (2) Tend to disagree (3) Disagree (4)
8. A test score of less than 50% clearly represents a very low level of achievement.
   Agree (1) Tend to agree (2) Tend to disagree (3) Disagree (4)
9. The minimum passing score on a test should be based at least in part on the scores earned by students of marginal ability who have been putting forth satisfactory effort.
   Agree (4) Tend to agree (3) Tend to disagree (2) Disagree (1)
10. The minimum passing score on a test should be based at least in part on how many of the test questions are easy enough for students of marginal ability to answer if they have been putting forth satisfactory effort. Agree (4) Tend to agree (3) Tend to disagree (2) Disagree (1)
11. Where to draw the line between grades of A and B, B and C, etc., should be decided only after looking at the distribution of scores for the class.
    Agree (4) Tend to agree (3) Tend to disagree (2) Disagree (1)
12. When a test turns out to be more difficult than the teacher intended, it is proper to "curve" the grades so that reasonable proportions of students receive As, Bs, Cs, etc.
    Agree (4) Tend to agree (3) Tend to disagree (2) Disagree (1)
13. A student's ability should be taken into consideration in awarding the final grade.
    Agree (1) Tend to agree (2) Tend to disagree (3) Disagree (4)
14. The amount of knowledge a student gains over the instructional period should be taken into consideration in awarding the final grade (in addition to the student's final level of achievement).
    Agree (1) Tend to agree (2) Tend to disagree (3) Disagree (4)
15. Laudatory or disruptive classroom behavior should be considered in determining final grades.
    Agree (1) Tend to agree (2) Tend to disagree (3) Disagree (4)
16. An exceptionally low or high degree of effort should be recognized by adjustment of the final grade.
    Agree (1) Tend to agree (2) Tend to disagree (3) Disagree (4)

Scale 1 Total 1-4 ↓

Scale 2 Total 5-8 ↓

Scale 3 Total 9-12 ↓

Scale 4 Total 13-16 ↓

For the remaining statements, assume that there is no school or district policy that dictates specific percentage ranges for grades.

17. Letter grades can be assigned to test scores more confidently if the test is hard enough to make the scores range from, say, 30% to 90%, rather than a range of 70% to 100%, as would result from an easier test. Agree (4) Tend to agree (3) Tend to disagree (2) Disagree (1)
18. Tests used to award letter grades should contain at least a moderate proportion of questions that fewer than half of the students can answer correctly.
    Agree (4) Tend to agree (3) Tend to disagree (2) Disagree (1)
19. In order to discriminate among the better students, it is desirable to use test questions that many of the less able students may be unable to answer correctly no matter how hard they have studied.
    Agree (4) Tend to agree (3) Tend to disagree (2) Disagree (1)
20. A harder test with an average score of 60% is fairer to students than an easier test with an average score of 80%, if the same numbers of A's, B's, C's, etc., will be assigned in either case.
    Agree (4) Tend to agree (3) Tend to disagree (2) Disagree (1)

Scale 5 Total 17-20 ↓

Now add the numbers in each set of four boxes. These are your scores on Scales 1 through 5. Write these totals on the corresponding lines below, fold under on the dotted line, and turn the page over.

Scale 1 ____ **Scale 2** ____ **Scale 3** ____ **Scale 4** ____ **Scale 5** ____

• Robert B. Frary, Lawrence H. Cross, and Larry J. Weber
Virginia Polytechnic Institue and State University
Blacksburg, VA 24061

Mark your scores on each vertical line of the figure below and to the right and connect the marks to reveal your score profile. Research has identified five patterns of scores across Scales 1-5 that identify groups of teachers with fairly distinct and well-defined opinions about testing and grading. However, a substantial proportion of teachers cannot be said to belong to any of the five groups. To see whether your scores suggest that you may be a member of one of these groups, follow the path indicated by the following statements:

1. If Scale 5 is 8 or less, go to 2. Otherwise go to 3.
2. If Scale 1 is 8 or less, go to 4. Otherwise go to 5.
3. If Scale 1 is 9 or less, go to 6. Otherwise go to 7.
4. If Scale 2 is 9 or less, go to 8. Otherwise go to 17.
5. If Scale 2 is 8 or less, go to 9. Otherwise go to 17.
6. If Scale 2 is 10 or less, go to 10. Otherwise go to 17.
7. If Scale 2 is 9 or less, go to 17. Otherwise go to 11.
8. If Scale 3 is 8 or less, go to 12. Otherwise go to 13.
9. If Scale 3 is 8 or less, go to 17. Otherwise go to 14.
10. If Scale 3 is 8 or less, go to 17. Otherwise go to 15.
11. If Scale 3 is 9 or less, go to 17. Otherwise go to 16.
12. If Scale 3 is 9 or less, go to 17. Otherwise go to 21.
13. If Scale 4 is 8 through 12, go to 18. Otherwise go to 17.
14. If Scale 4 is 10 or less, go to 19. Otherwise go to 17.
15. If Scale 4 is 9 or less, go to 20. Otherwise go to 17.
16. If Scale 4 is 8 or less, go to 17. Otherwise go to 22.
17. Your scores do not place you in any of the five groups. Compare your score profile to those described below. If you miss a group by one point on one scale you may belong to it.
18. Your score profile is within the shaded area of Figure 1. This group tends to favor districtwide percentage grading scales and to believe that percent-correct scores reveal absolute levels of knowledge. This strict interpretation is alleviated by use of relatively easy tests, a willingness to adjust cut points for grades, and some tendency to adjust grades on the basis of nonachievement criteria. While fundamentally inconsistent, this approach to testing and grading may result simply from a desire to avoid too many low grades. Use of nonachievement factors mainly to raise rather than lower grades would be compatible with this conjecture.
19. Your score profile is within the shaded area of Figure 2. This group tends to disapprove of districtwide percentage scales and not to believe that percent-correct scores reveal absolute levels of knowledge. Consistent with these views they are often willing to vary score cut points for grades. However, they also often use nonachievement criteria in determining course grades and tend to prefer easy tests. Some of these teachers might be viewed as softhearted. Others may be antitesting.
20. Your score profile is within the shaded area of Figure 3. This group tends to approve of district-wide percentage grading scales and to believe that percent-correct scores reflect absolute levels of knowledge. Inconsistent with these positions, they tend toward flexibility in adjusting cut points for grades. Combining these characteristics with rather heavy use of nonachievement criteria in determination of course grades and advocacy of relatively difficult tests suggests a possibly arbitrary or manipulative approach to grading.
21. Your score profile is within the shaded area of Figure 4. This group tends to approve of district-wide percentage grading scales and to believe that percent-correct scores reflect absolute levels of knowledge. Consistent with these positions, they also tend to reject flexibility in determining cut points for grade assignment and the use of nonachievement factors in determining course grades. Such a stern or "hardnosed" approach to testing and grading could result in large numbers of very low grades, but this group tends to avoid this outcome through the use of relatively easy tests.
22. Your score profile is within the shaded area of Figure 5. This is the only group generally in agreement with what most measurement specialists would recommend. Its members tend to reject districtwide percentage grading scales, to understand that percent-correct scores do not directly reflect a student's level of knowledge, to favor difficult tests (that maximize ranking accuracy), to approve of flexibility (based on test characteristics) in determining cut scores for grades, and to reject the use of nonachievement factors in determining course grades.